W38–W42 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm224

Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics

Joaquín Tárraga[1,2], Ignacio Medina[1], Leonardo Arbiza[1], Jaime Huerta-Cepas[1,2], Toni Gabaldón[1], Joaquín Dopazo[1,2] and Hernán Dopazo[1,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF) and [2]Functional Genomics Node, INB, CIPF, Valencia 46013, Spain

Received January 30, 2007; Revised March 23, 2007; Accepted March 28, 2007

ABSTRACT

Phylemon is an online platform for phylogenetic and evolutionary analyses of molecular sequence data. It has been developed as a web server that integrates a suite of different tools selected among the most popular stand-alone programs in phylogenetic and evolutionary analysis. It has been conceived as a natural response to the increasing demand of data analysis of many experimental scientists wishing to add a molecular evolution and phylogenetics insight into their research. Tools included in Phylemon cover a wide yet selected range of programs: from the most basic for multiple sequence alignment to elaborate statistical methods of phylogenetic reconstruction including methods for evolutionary rates analyses and molecular adaptation. Phylemon has several features that differentiates it from other resources: (i) It offers an integrated environment that enables the direct concatenation of evolutionary analyses, the storage of results and handles required data format conversions, (ii) Once an outfile is produced, Phylemon suggests the next possible analyses, thus guiding the user and facilitating the integration of multi-step analyses, and (iii) users can define and save complete pipelines for specific phylogenetic analysis to be automatically used on many genes in subsequent sessions or multiple genes in a single session (phylogenomics). The Phylemon web server is available at http://phylemon.bioinfo.cipf.es.

INTRODUCTION

Phylogenetic and evolutionary analyses of sequences are among the most often used methodologies in laboratories working in functional, comparative and structural genomics (1). Since 1980, when the first version of the PHYLogeny Inference Package (PHYLIP) (2) was introduced by Felsenstein, a high number of programs for phylogenetic inference have been developed. Currently, PHYLIP(2), PAUP*(3), MEGA(1), PhyML(4), PAML(5) and MrBayes(6) are well-known programs that are used by thousands of users around the world. Other more specific programs, designed to test evolutionary hypotheses for model selection, tree topology, molecular clock or adaptation, are less popular among common users, but they are, nevertheless, of great interest for users familiar with evolutionary enquiries. Currently, the most comprehensive list of phylogenetic resources can be found at the University of Washington in Seattle (http://evolution.genetics.washington.edu/phylip/software.html), which listed 292 phylogeny packages and 38 web servers, by 2003.

Web servers for phylogenetic and evolutionary analyses provide a direct means for addressing several evolutionary questions, ranging from the computation of a multiple alignment and a neighbor-joining tree using ClustalW program (7) (http://www.ebi.ac.uk/clustalw/), to the more sophisticate analysis of molecular adaptation for detection of positively selected sites in DNA sequences (using methods as those available in the HYPHY package (8) (http://www.datamonkey.org/)). Many such servers run a single tool or program whereas others bring together many of the most popular programs of phylogenetic reconstruction (e.g. see http://bioweb.pasteur.fr/seqanal/phylogeny/intro-uk.html).

Despite this diversity there is, so far, no single integrated web server that provides a common framework to run the most frequent analyses on DNA and protein sequences from a phylogenetic and evolutionary perspective. Non-expert users are then often overwhelmed by the variety of servers, formats and options available and by the difficulty of concatenating analyses performed on different servers. The main objective of Phylemon is to fulfil this need by providing users with the possibility of finding all the necessary applications in a single integrated web framework that guides them throughout the whole evolutionary analysis.

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: hdopazo@cipf.es

OUTLINE OF THE PROGRAM

Phylemon is a web server that integrates a selected suite of more than 20 different tools from the most popular stand-alone programs of phylogenetic and evolutionary analysis (Figure 1A).

Three features characterize all tools integrated in Phylemon: (1) tools have available examples in order to familiarize users with the correct input data and expected results, (2) input formats (preferentially FASTA or PHYLIP) are automatically transformed in order to move among alternative tools and (3) all the input and output result files can be saved in default or user-defined projects (folders).

Phylemon can be accessed by anonymous login or by registered users. The only difference between these choices is that registered users, from whom only an e-mail is required, can store project results and use them at a later time for further analysis (Figure 1G).

PHYLOGENETIC PROGRAMS

Phylemon runs distance-based methods, maximum parsimony analyses and statistical methods of phylogenetic reconstruction. Distances and parsimony methods for DNA or protein sequence data are provided by the most often used algorithms of the PHYLIP package (2) v3.65: DnaDist, ProtDist, DnaPars and ProtPars, respectively. Pairwise distance matrices can be represented in a phylogenetic tree using the neighbor-joining (NJ) algorithm (Neighbor) or applying a least square (LS) method or a minimum evolution (ME) criterium (Fitch program). In order to obtain trees with statistical support on internal nodes, a re-sampling method (i.e. bootstrap option included in the Seqboot algorithm) and the corresponding trees summarizing algorithm (i.e. majority rule tree using Consense program) of PHYLIP are included.

Basic maximum likelihood (ML) analyses of DNA and protein sequence data are provided with the DnaML and ProML algorithms of the PHYLIP package in Phylemon. When a more sophisticated ML analysis is required users can run PhyML version aLRT (9,10) or TREE-PUZZLE v5.2 (11). Major differences between these ML programs are: (1) PhyML is faster than any other ML algorithm of phylogenetic reconstruction, (2) TREE-PUZZLE uses a quartet-puzzling method instead the more classical heuristic searches for tree searching, (3) TREE-PUZZLE reports reliability values while the PhyML method reports Felsenstein's bootstrap values and aLRT-related statistics branch support (9), (4) TREE-PUZZLE can quantify the amount of the phylogenetic signal contained in a data set (the probability of the data producing a tree-like phylogenetic representation) through the likelihood-mapping method (12) and (5) TREE-PUZZLE computes ML pairwise distances that can easily be represented in an NJ/LS/ME tree.

Finally, Phylemon runs Bayesian phylogenetic analysis using MrBayes v3.1.12 (6). MrBayes runs in Phylemon with the same characteristics that users have in Windows or Linux interfaces. Users can define all the parameters of MrBayes in a file to upload to the server or, alternatively, edit the parameters in the specific text box when entering the data. A valuable list of examples showing alternative data and parameters is available in the server. In addition, when users edit sequence data or parameters, the MrBayes command list is available on the fly. At the end of the run, Phylemon asks for the sump and sumt parameters in order to define the burnin limit (Figure 1D). Once the analysis is finished, 10 outfiles are listed, 8 corresponding to the usual files produced by MrBayes and 2 corresponding to: (1) all information printed to the standard output the program runs (outfile.out) and (2) the last topologies retained (with branch lengths and posterior probabilities on the nodes—tree.nw), the last of which can be visualized in a tree-viewer program available in the utilities section of Phylemon.

TESTING MODELS, TOPOLOGIES AND CLOCKS

Evolutionary biologists frequently use likelihood ratios or alternative statistical test to deal with hypotheses about phylogenies or evolutionary models. Users can run three kinds of tests in Phylemon: (1) tests on models of evolution [ModelTest (13), ProtTest (14)], including tests on molecular adaptation [using PAML v3.15 (5) and SLR (15) programs], (2) tests on topologies [paired-sites test (16) using ML programs of PHYLIP and TREE-PUZZLE] and (3) tests on molecular clocks [relative rate test using RRTree (17)].

MODEL SELECTION

Selecting the best-fit model of evolution for a data set can be done using the statistical methodology implemented in ModelTest and ProtTest for DNA and protein sequences, respectively. ModelTest and ProtTest run in Phylemon under the HyPhy (Hypothesis Testing Using Phylogenies) environment (8). That means that likelihoods, hierarchical LRT (hLRTs) comparisons and the Akaike information criterion (AIC) are computed without requiring any other program. Although ModelTest in HyPhy runs all the original DNA models, ProtTest in HyPhy manages half of the models included in the original ProtTest. Nevertheless, the most general amino acid models are included in the Phylemon server.

MOLECULAR ADAPTATION

Molecular adaptation is an exciting topic in molecular evolution and phylogenetic studies (18). Three alternative programs are included in Phylemon that allow the inference of molecular adaptation events, these are YN00 and CodeML from PAML v3.15 (5) and the sitewise likelihood-ratio (SLR) (15) method.

YN00 program implements pairwise computations of ω (dN/dS) from synonymous and non-synonymous substitutions rates as defined in different counting methods (18), such as NG (19), LWL (20), Li (21), PB (22) and YN00 (23) (Figure 1F).

CodeML uses numerical optimization algorithms to maximize the log-likelihood values under a specific model of evolution. CodeML requires that users provide option

parameters in a control file in which all variables of the ML models are listed. Although likely straightforward for advanced users, the configuration and compatibility of the different options are not evident for novel users. Therefore, we have developed a web interface with more than 20 examples covering the branch, site and branch-site ML models. Branch models allow searching for positive selection acting on a particular lineage in a phylogeny (24), whereas site models detect adaptive evolution on codon positions in the alignment (25,26), and branch-site models detect positive selection affecting only a few sites along a few lineages (27,28) (see (29,30) for its application on the human genome).

The SLR method uses a site by site approach to test for neutrality but, in contrast to similar methods such as SG (31), SLR does so by using the entire alignment to determine quantities common to all sites, such as evolutionary distances. At the end of the test, a necessary correction for multiple testing is completed. Readers can see (32) for a comparison on estimations of SLR and CodeML site models.

RELATIVE RATES TEST

Substitution rates between DNA or protein sequences, whether grouped or not in phylogenetically defined lineages, can be statistically compared in Phylemon. This is done by using relative rates test (33) as computed in RRTree vs 1.1.11 (17,34) program. RRTree computes relative rates tests among user-defined lineages. When a lineage includes many species, RRTree computes relative rates tests with a weighted scheme for species based on the tree topology provided by the user (34,35).

RRTree computes differences in rates for coding DNA sequences using different parameters: the number of synonymous substitutions and synonymous transitions per synonymous site (Ks and As, respectively), the number of non-synonymous substitutions and non-synonymous transversions per non-synonymous site (Ka and Ba, respectively) and, finally, the number of synonymous transversions per 4-fold degenerate site (B4). Kimura two parameters (K2P) (36) and Jukes and Cantor (JC) models are available for non-coding DNA sequences. For protein sequences, RRTree computes a modification of JC model (17).

TOPOLOGIES TEST

Sometimes it is interesting to test among different alternatives which topology best explain a particular data set. The basic idea of paired-sites tests is that two trees can compared through either their parsimony or likelihood scores computed from the distribution of the costs or likelihoods for each site (see Chapter 21 of (2) for a detailed description of these tests). Paired-sites test using ML can be performed using DnaML, ProML (from PHYLIP package) by means of the KH test (37) or using TREE-PUZZLE for SH (38) and ELW (39) tests (Figure 1E). In any case, the option of evaluation of user-defined trees must be selected form the tree search option dialog. The topology that best fit the data, according to a pre-specified model of sequence evolution, is presented at the end of the run.

PIPELINES AND PHYLOGENOMICS

Phylogenomic analyses sometimes involve repeating a certain set of analysis over several orthologous groups of genes. In such cases, it is necessary to apply common phylogenetic algorithms to different sequence data using a single pipeline of tools. For instance: ClustalW, Seqboot, DnaDist/ProtDist, Neighbor and Consense may be used in that order for a phylogenetic reconstruction with bootstrap values. We have developed a Java applet environment (the Super-Phylemon Pipeliner) in order to satisfy this requirement (Figure 1B). Users interested in such kind of studies can upload all the files containing the sequences (up to a maximum of 20) in order to run one or more pipelines. Currently, Super-Phylemon provides basic programs derived from the PHYLIP package. Future versions of Super-Phylemon will include all the tools of phylogenetics and evolutionary tests included in Phylemon in order to facilitate the implementation of more complex pipelines.

OTHERS TOOLS AND UTILITIES

The Phylemon web server integrates two different programs for the alignment of multiple sequences: ClustalW v1.83 (7) and MUSCLE v3.52 (40). Furthermore, Phylemon provides additional pre- and post-analysis utilities. These include file format conversion, gene concatenation, tree visualization and the computation of distances between trees.

Conversion between sequence formats can be made by means of the ReadSeq program (GNU/Linux program). Users can transform alternative files to FASTA or PHYLIP format and run with confidence any of the Phylemon tools. The concatenation of individual multiple alignments (PHYLIP format) with equal or different number of species, generally employed in phylogenomic studies, can be made using a facility specifically developed by us for such a purpose.

Rooted and unrooted newick tree formats can be visualized in rectangular, radial and circular diagrams using ETE program (Environment for Tree Exploration, developed by JHC, (Figure 1C)). Finally topological distances between trees are computed by the TreeDist program from PHYLIP. This program measures symmetric differences or branch score distances between two or more trees [see (41) for an application of its use].

DISCUSSION

Molecular evolution and phylogenetics embrace a wide range of scientific enquiries. Following the development of the field in the last 20 years, researchers have developed tools ranging from the most complete packages to the more specific programs. Although some of these are available online in separate, dedicated web servers, many of the programs available in Phylemon cannot be found on any public web server. This is the case for the

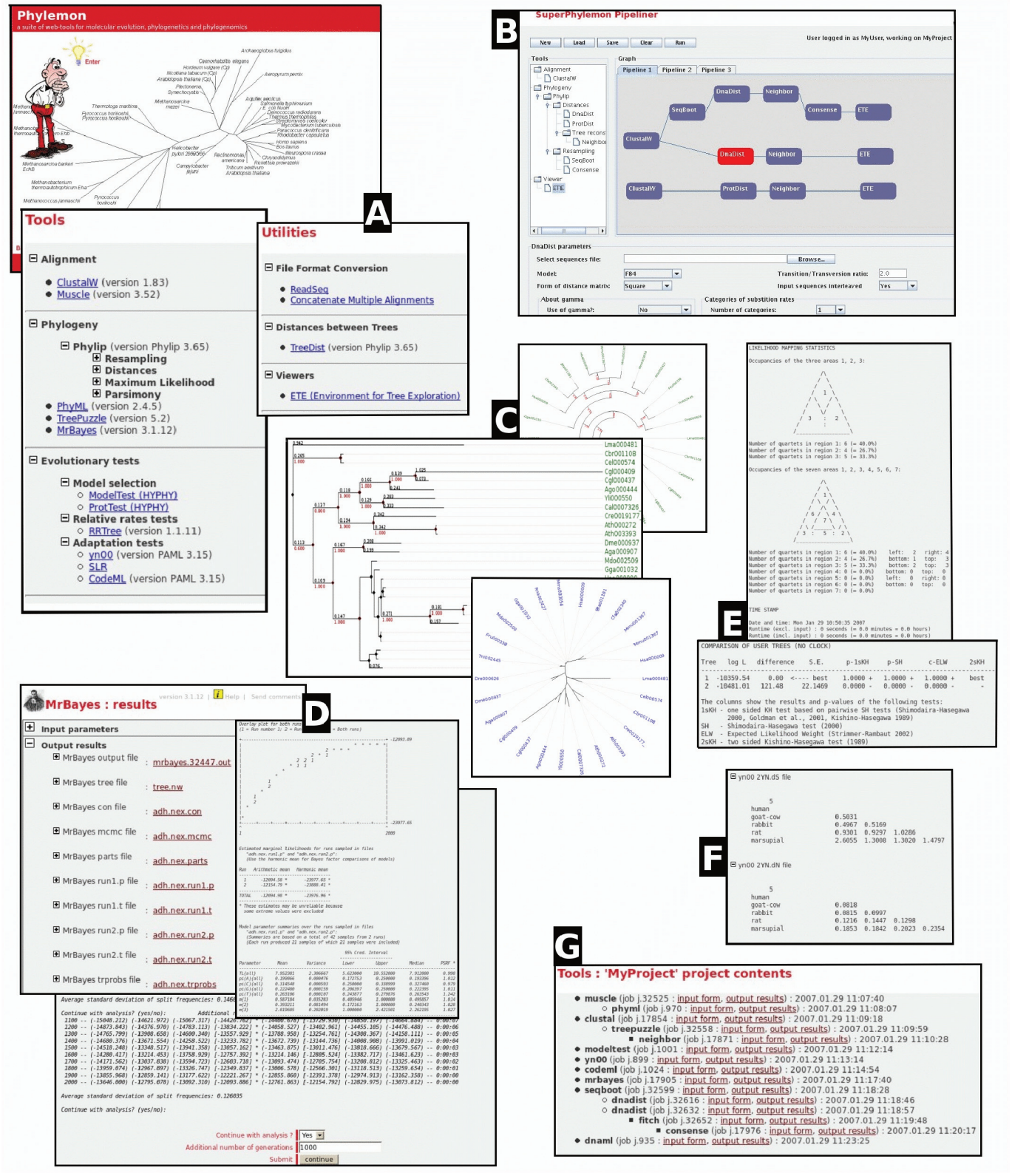

Figure 1. Phylemon tools and selected analyses. (**A**) Home page, tools and utilities of Phylemon web server. (**B**) SuperPhylemon java applet showing a user-defined pipeline (among three) with three alternative analyses coming from two ClustalW files. Program's parameters are setting below the pipeline (for instance, DnaDist-red box). (**C**) Rectangular, circular and radial trees drawing with ETE (Environment for Tree Exploration) program. (**D**) MrBayes analysis showing the output files, the likelihood increase diagram and the box dialog for additional MCMC generations. (**E**) Likelihood-mapping and topology-test results from TREE-PUZZLE. (**F**) A detailed portion of YN00 results showing dS and dN pairwise computation among five species. (**G**) Registered users (MyUser, see B) can save input and output files in specific projects (MyProject) of Phylemon web server.

frequently used MrBayes, Tree-Puzzle, CodeML in full, SLR and RRTree programs.

Altogether, Phylemon addresses an important, yet unanswered, necessity of users working with evolutionary and phylogenetic analysis of molecular sequences; namely, the need for a public web server providing a core set of format compatible tools truly integrated in an independent platform.

ACKNOWLEDGEMENTS

This work is supported by grants from Generalitat Valenciana GV06/080 and MEC BFU2006-15413-C02-02/BMC to H.D., J.T., J.H.C. and Functional Genomics node (INB) are funded by Genoma España. T.G. is recipient of an EMBO postdoctoral fellowship LTF 402-2005. Funding to pay the Open Access publication charges for this article was provided by MEC project to HD.

Conflict of interest statement. None declared.

REFERENCES

1. Kumar,S., Tamura,K. and Nei,M. (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. *Comput. Appl. Biosci.*, **10**, 189–191.
2. Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) Version 3.6. Distributed by the Author*. Department of Genome Sciences, University of Washington, Seattle.
3. Swofford,D.L. (2002). In: Associates,S. (ed.), *PAUP* phylogenetic Analysis using parsimony (*and other Methods)*. Version 4. Sinauer Associates, Sunderland, Massachusetts, USA.
4. Guindon,S. and Gascuel,O. (2003) PhyML – A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol., **52**, 696–704.
5. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
6. Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
7. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
8. Pond,S.L., Frost,S.D. and Muse,S.V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.
9. Anisimova,M. and Gascuel,O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.*, **55**, 539–552.
10. Guindon,S., Lethiec,F., Duroux,P. and Gascuel,O. (2005) PHYML Online – a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.*, **33**, W557–W559.
11. Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
12. Strimmer,K. and von Haeseler,A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. USA*, **94**, 6815–6819.
13. Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
14. Abascal,F., Zardoya,R. and Posada,D. (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.
15. Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.
16. Felsenstein,J. (2004) *Inferring Phylogenies* Sinauer Associates, Sunderland, MA, USA.
17. Robinson-Rechavi,M. and Huchon,D. (2000) RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics*, **16**, 296–297.
18. Yang,Z. (2006) *Computational Molecular Evoltion* Oxford University Press, Oxford, UK.
19. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
20. Li,W.H., Wu,C.I. and Luo,C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.
21. Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.
22. Pamilo,P. and Bianchi,N.O. (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.
23. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
24. Yang,Z. and Nielsen,R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.*, **46**, 409–418.
25. Nielsen,R. and Yang,Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.
26. Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.
27. Yang,Z. and Nielsen,R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.
28. Zhang,J., Nielsen,R. and Yang,Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, **22**, 2472–2479.
29. Clark,A.G., Glanowski,S., Nielsen,R., Thomas,P.D., Kejariwal,A., Todd,M.A., Tanenbaum,D.M., Civello,D., Lu,F. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
30. Arbiza,L., Dopazo,J. and Dopazo,H. (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.*, **2**, e38.
31. Suzuki,Y. and Gojobori,T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, **16**, 1315–1328.
32. Arbiza,L., Duchi,S., Montaner,D., Burguet,J., Pantoja-Uceda,D., Pineda-Lucena,A., Dopazo,J. and Dopazo,H. (2006) Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J. Mol. Biol.*, **358**, 1390–1404.
33. Sarich,V.M. and Wilson,A.C. (1973) Generation time and genomic evoltion in primates. *Science*, **179**, 1144–1149.
34. Robinson,M., Gouy,M., Gautier,C. and Mouchiroud,D. (1998) Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.*, **15**, 1091–1098.
35. Dopazo,H. and Dopazo,J. (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.*, **6**, R41.
36. Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.
37. Kishino,H. and Hasegawa,M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.*, **29**, 170–179.
38. Shimoaira,H. and Hasegawa,M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, **16**, 1114–1116.
39. Strimmer,K. and Rambaut,A. (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.*, **269**, 137–142.
40. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
41. Dopazo,H., Santoyo,J. and Dopazo,J. (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics*, **20**(Suppl. 1), I116–I121.